Hotel Reservations Classification

Zenthia Song

Overview

- TASK
  - I want to train a binary classification model to determine if each customer is likely to cancel their reservation or not.
- DATASET
  - https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset?datasetId=2783627&sortBy=voteCount
- METHOD
  - Logistic Regression, KNeighbours, Decision Tree, and Random Forest.

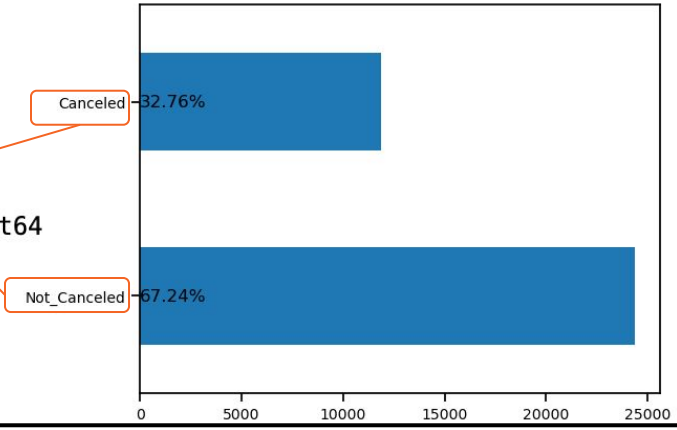
Summary Data

```
Booking_ID                            object
no_of_adults                           int64
no_of_children                         int64
no_of_weekend_nights                   int64
no_of_week_nights                      int64
type_of_meal_plan                     object
required_car_parking_space             int64
room_type_reserved                    object
lead_time                              int64
arrival_year                           int64
arrival_month                          int64
arrival_date                           int64
market_segment_type                   object
repeated_guest                         int64
no_of_previous_cancellations           int64
no_of_previous_bookings_not_canceled   int64
avg_price_per_room                   float64
no_of_special_requests                 int64
booking_status                        object
dtype: object
```

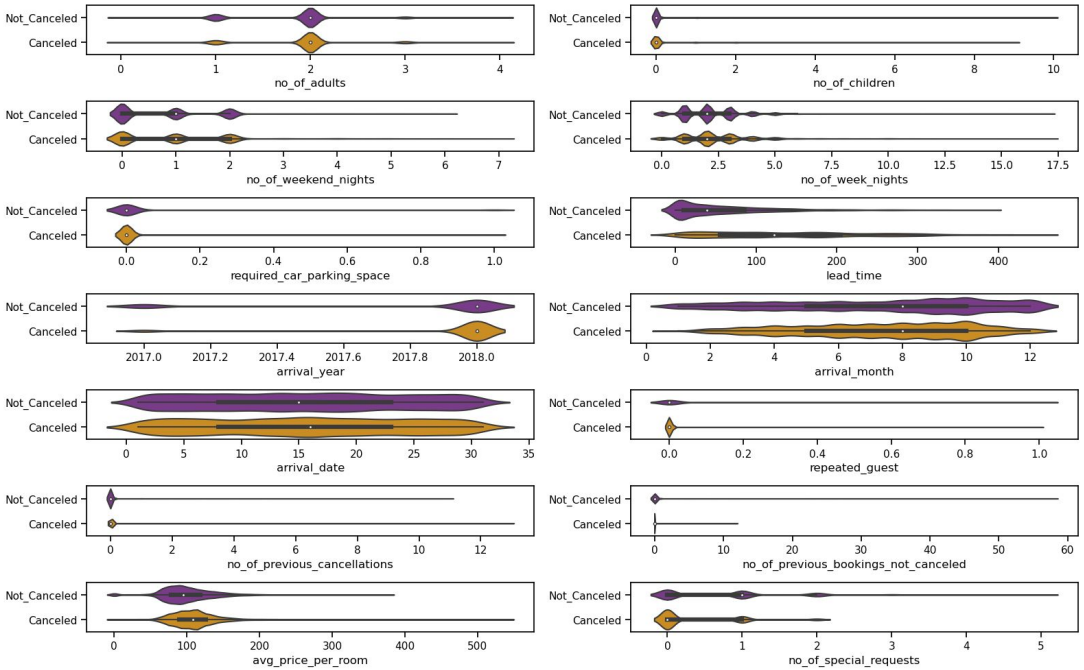
	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arriv
0	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	
1	INN00002	2	0	2	3	Not Selected	0	Room_Type 1	5	
2	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	
3	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	
4	INN00005	2	0	1	1	Not Selected	0	Room_Type 1	48	
...	...	...	...	...	...	...	...	...	...	...
36270	INN36271	3	0	2	6	Meal Plan 1	0	Room_Type 4	85	
36271	INN36272	2	0	1	3	Meal Plan 1	0	Room_Type 1	228	
36272	INN36273	2	0	2	6	Meal Plan 1	0	Room_Type 1	148	
36273	INN36274	2	0	0	3	Not Selected	0	Room_Type 1	63	
36274	INN36275	2	0	1	2	Meal Plan 1	0	Room_Type 1	207	

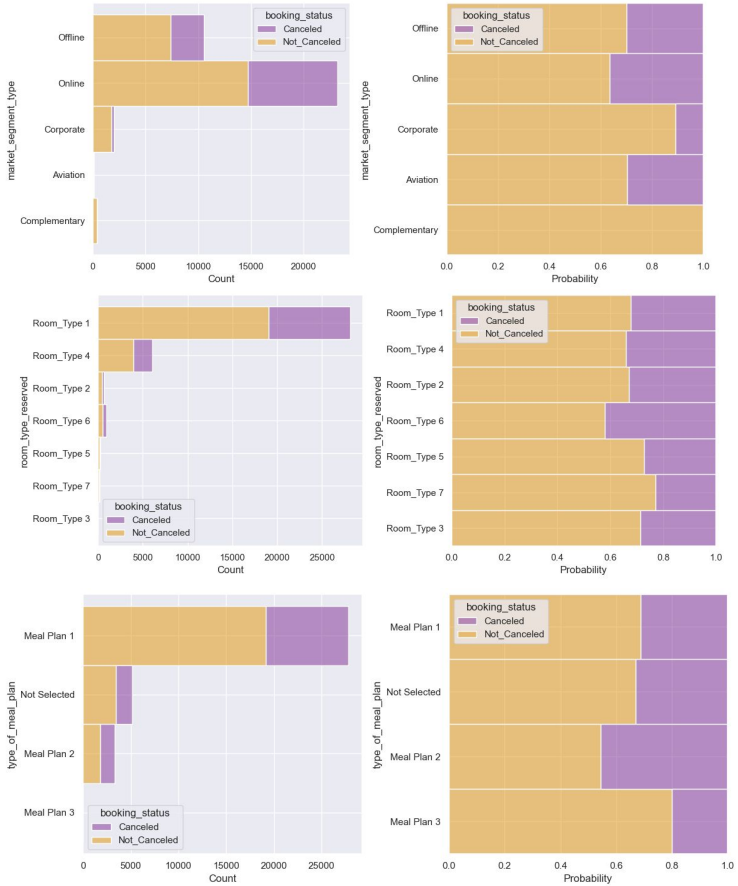
36275 rows × 19 columns

```
Not_Canceled    24390
Canceled        11885
Name: booking_status, dtype: int64
```

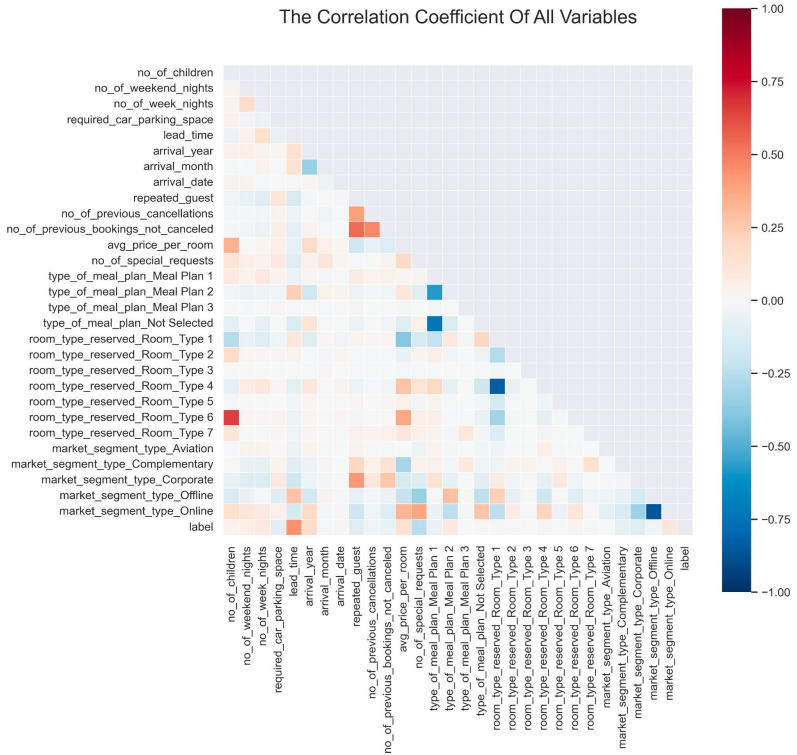


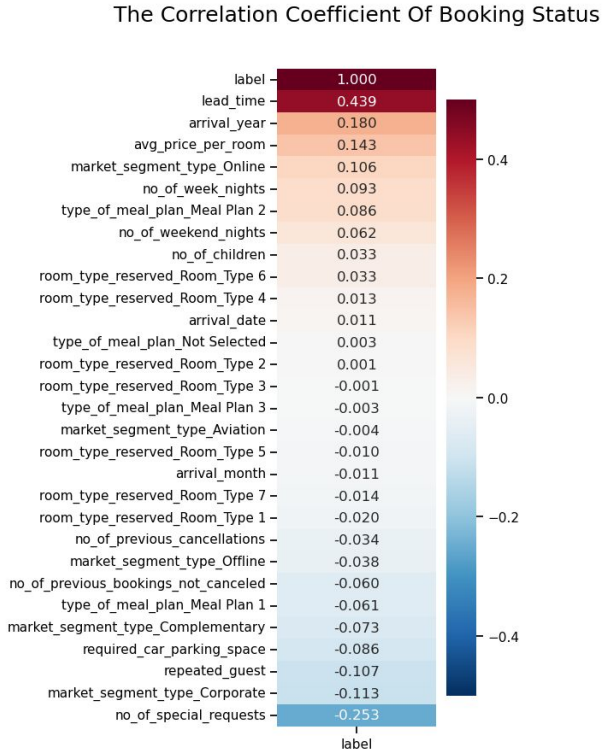
Summary Data



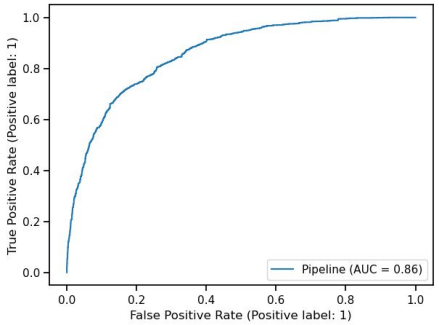


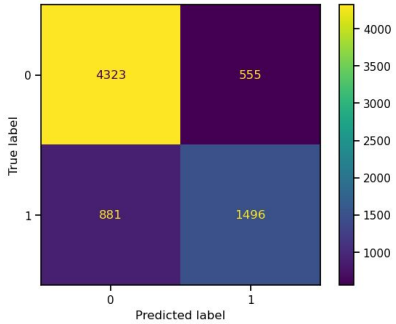
Correlation Visualization





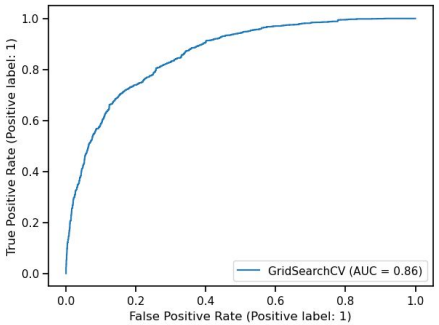
Model: Logistic Regression

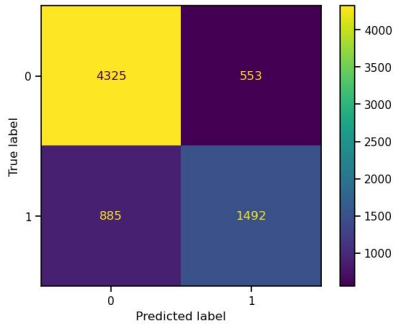




	precision	recall	f1-score	support
0	0.83	0.89	0.86	4878
1	0.73	0.63	0.67	2377
accuracy			0.80	7255
macro avg	0.78	0.76	0.77	7255
weighted avg	0.80	0.80	0.80	7255

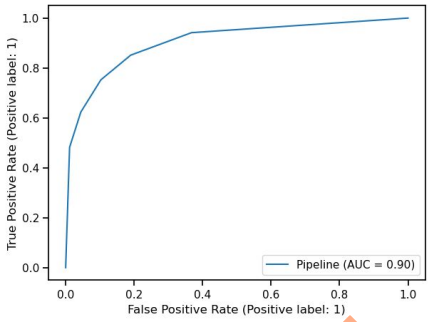
Default Parameters vs After Parameters Search

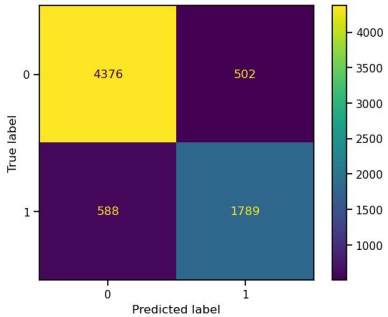




	precision	recall	f1-score	support
0	0.83	0.89	0.86	4878
1	0.73	0.63	0.67	2377
accuracy			0.80	7255
macro avg	0.78	0.76	0.77	7255
weighted avg	0.80	0.80	0.80	7255

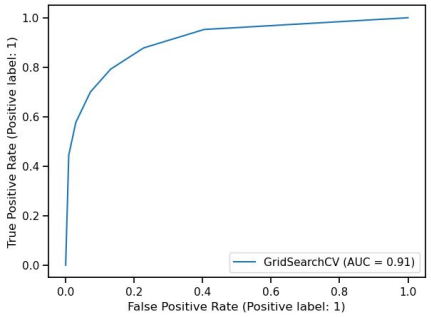
Model: K Neighbors

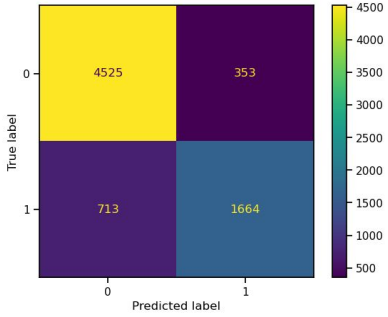




	precision	recall	f1-score	support
0	0.88	0.90	0.89	4878
1	0.78	0.75	0.77	2377
accuracy			0.85	7255
macro avg	0.83	0.82	0.83	7255
weighted avg	0.85	0.85	0.85	7255

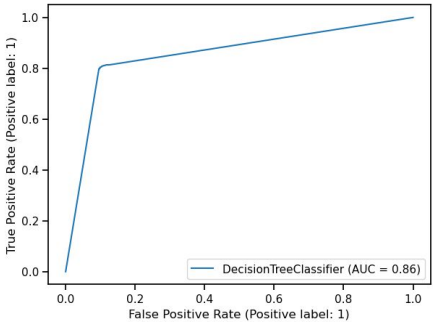
Default Parameters vs After Parameters Search

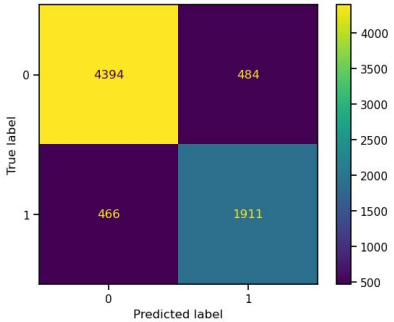




	precision	recall	f1-score	support
0	0.86	0.93	0.89	4878
1	0.82	0.70	0.76	2377
accuracy			0.85	7255
macro avg	0.84	0.81	0.83	7255
weighted avg	0.85	0.85	0.85	7255

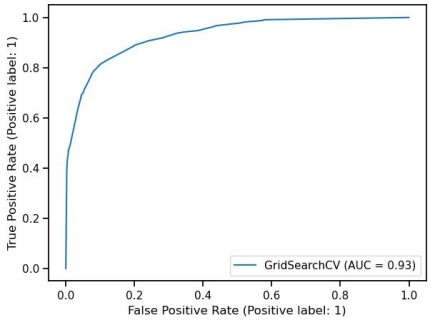
Model: Decision Tree

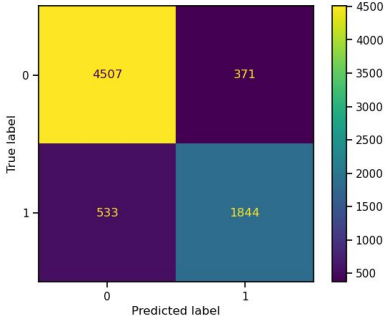




	precision	recall	f1-score	support
0	0.90	0.90	0.90	4878
1	0.80	0.80	0.80	2377
accuracy			0.87	7255
macro avg	0.85	0.85	0.85	7255
weighted avg	0.87	0.87	0.87	7255

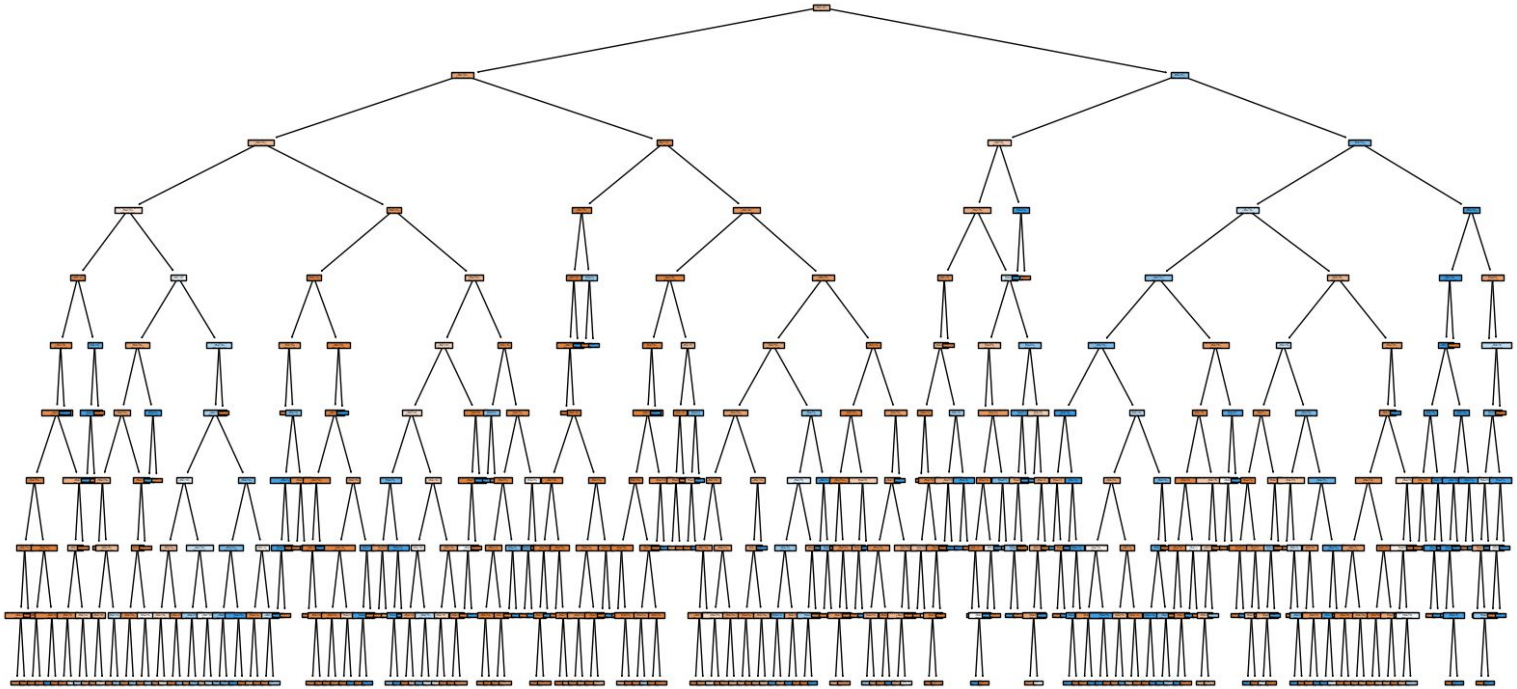
Default Parameters vs After Parameters Search



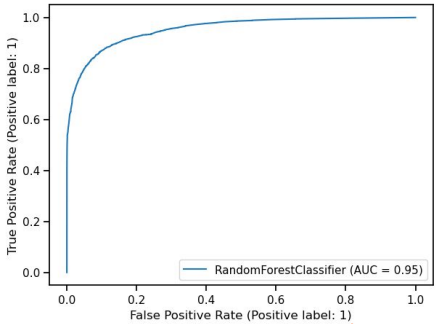


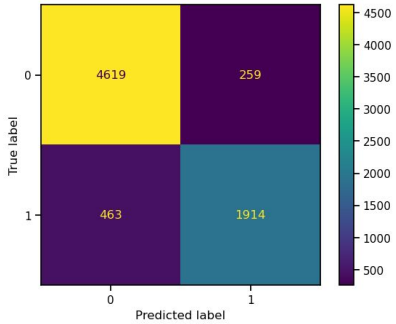
	precision	recall	f1-score	support
0	0.89	0.92	0.91	4878
1	0.83	0.78	0.80	2377
accuracy			0.88	7255
macro avg	0.86	0.85	0.86	7255
weighted avg	0.87	0.88	0.87	7255

Model: Decision Tree



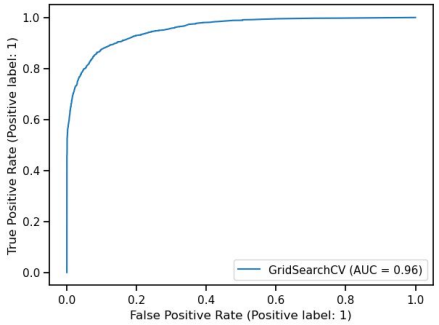
Model: Random Forest

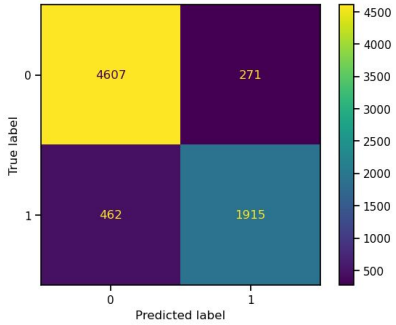




	precision	recall	f1-score	support
0	0.91	0.94	0.93	4878
1	0.88	0.81	0.84	2377
accuracy			0.90	7255
macro avg	0.89	0.88	0.88	7255
weighted avg	0.90	0.90	0.90	7255

Default Parameters vs After Parameters Search





	precision	recall	f1-score	support
0	0.91	0.94	0.93	4878
1	0.88	0.81	0.84	2377
accuracy			0.90	7255
macro avg	0.89	0.88	0.88	7255
weighted avg	0.90	0.90	0.90	7255

Comparing Model Performance

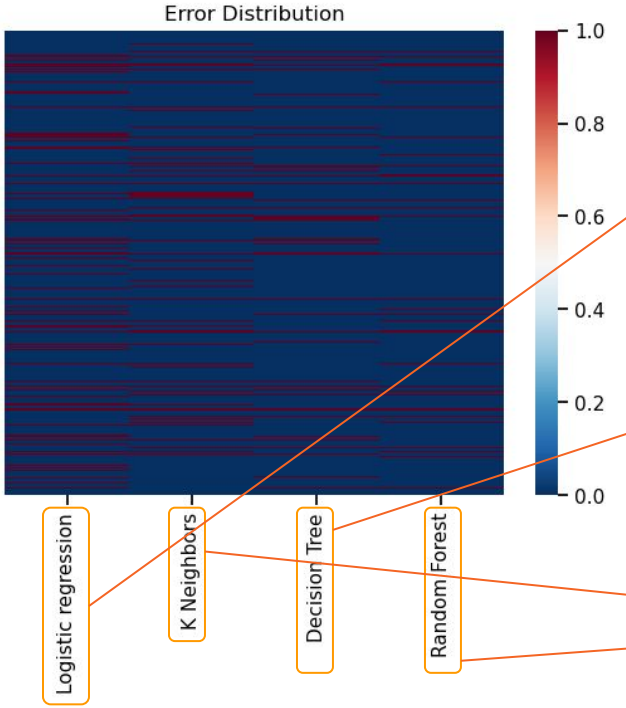
The Random Forest model

- ➔ achieved the highest AUC score of 0.96, indicating that it has a high ability to distinguish between positive and negative classes.
- ➔ achieved the highest accuracy score of 0.90, indicating that it correctly classified 90% of instances.

Overall, the Random Forest model achieved the highest scores in both AUC and accuracy, indicating that it is the best model for this dataset.

	Model	AUC	Accuracy	Weighted avg Precision	Weighted avg Recall	weighted avg F1
3	Random Forest	0.96	0.90	0.90	0.90	0.90
2	Decision Tree	0.93	0.88	0.87	0.87	0.87
1	K Neighbors	0.90	0.85	0.85	0.85	0.85
0	Logistic regression	0.86	0.80	0.80	0.80	0.80

Error Analysis



	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time
20514	0	0	2	0	192
32203	0	2	3	0	212
25059	0	0	2	0	103
5249	0	1	3	0	106
19569	0	1	3	0	50
...	...	...	...	...	...
17530	0	1	2	0	81
29306	0	1	1	0	0
7287	0	2	1	0	92
13295	0	2	4	0	52
24803	0	0	2	0	192

1438 rows × 29 columns

	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time
5249	0	1	3	0	106
19569	0	1	3	0	50
35867	0	0	2	0	39
11247	0	0	1	0	3
12472	1	0	1	0	10
...	...	...	...	...	...
21684	0	2	1	0	96
17530	0	1	2	0	81
29306	0	1	1	0	0
7287	0	2	1	0	92
13295	0	2	4	0	52

1001 rows × 29 columns

	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time
5249	0	1	3	0	106
19569	0	1	3	0	50
35867	0	0	2	0	39
11247	0	0	1	0	3
35851	0	1	3	0	8
...	...	...	...	...	...
35366	0	0	1	0	0
2989	0	0	1	0	49
21684	0	2	1	0	96
17530	0	1	2	0	81
13295	0	2	4	0	52

1066 rows × 29 columns

	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time
32203	0	2	3	0	212
5249	0	1	3	0	106
19569	0	1	3	0	50
35867	0	0	2	0	39
11247	0	0	1	0	3
...	...	...	...	...	...
9493	0	1	1	0	64
2989	0	0	1	0	49
21684	0	2	1	0	96
2412	0	0	3	0	103
13295	0	2	4	0	52

733 rows × 29 columns

Error Analysis

Possible reasons for misprediction:

- Noise: Random noise in the data can have a negative impact on the model's performance.
- Bias: Bias refers to the model's deviation from the true relationship, usually due to the model being too simple or the assumptions being incorrect.
- Variance: Variance refers to the instability of the model's predictions on different datasets, usually due to the model being too complex or overfitting.
    - For example:
        - Logistic regression models can overfit the training data if the model is too complex or the regularization parameter is set too low.
        - Trees can easily overfit the training data if the tree is too deep or if the stopping criterion is not set correctly.
- Insufficient Features: If there are not enough features or the selected features are not relevant to the problem, the model's performance may also be affected.
- Imbalanced classes/data: it can lead to poor performance in rare classes.
    - For example:
        - KNN can be sensitive to imbalanced classes because it relies on the majority class in the k nearest neighbors.
        - Random Forest can be biased toward the majority class in imbalanced datasets.

To improve the performance of the model, the following methods can be used:

- Use more and better features to capture more patterns and relationships.
- Choose a model that is more suitable for the problem and avoid models that are either too simple or too complex.
- Use ensemble learning methods, such as gradient boosting trees, to improve the performance and robustness of the model.